基于文本挖掘的游客对古镇旅游态度的分析

范珈瑜
山东财经大学, 山东 济南 250014

摘要
围绕游客的反馈，分析古镇旅游项目存在的问题及游客的态度。选取了5个各具特色的古镇——甲居藏寨、西江千户苗寨、周庄古镇、黄姚古镇、西塘古镇作为参考。利用大量的评论进行文本挖掘、统计词频，分析热门话题。再运用情感分析得到评分，进而分析游客的好感度及可能影响评价的因素。最后，结合因子分析建立综合模糊评分模型，以计算古镇旅游的最终得分。研究结果可作为参考，以此优化改善古镇旅游体验，保护历史文明遗址。

关键词
文本挖掘；古镇旅游；情感分析；因子分析；模糊综合评价模型；方差分析

中图分类号：TP391 文献标识码：A doi: 10.11959/j.issn.2096-0271.2017064

Analysis of tourists' attitude for ancient towns based on text mining

FAN Jiayu
Shandong University of Finance and Economics, Jinan 250014, China

Abstract
The tourists' attitudes and the potential problems coming along with tourism projects were mainly analyzed. Thousands of comments aiming at five different ancient China towns coming from website were collected. Then with text mining, terms frequency was counted, and the hot topics were analyzed. After that, the scores were received with sentiment analysis. Based on the results from sentiment analysis, then the analysis of variance about possible causes influenced travelers' attitudes was given. Ultimately the fuzzy comprehensive evaluation model was built. The result is able to give references for creating a better system of ancient towns traveling program.

Key words
text mining, traveling of ancient towns, sentiment analysis, factor analysis, fuzzy comprehensive evaluation model, analysis of variance

1 引言

据中国旅游研究院发布的《中国旅游经济蓝皮书》和统计数据显示，中国旅游业飞速发展，2015年入境旅游3年来首次增长。旅游产业成为带动经济发展的重要驱动力。开发特色景区成为了许多经济滞后的偏远地区刺激经济发展的重要策略。其中，开发具有本土文化特色的古镇旅游很受欢迎。但是效益驱动的景区开发愈演愈烈，市场各种形式的旅游消费方式层出不穷。由于缺乏对旅游市场的管理经验和过度追逐经济效益等因素，导致许多景区开发过度，甚至市场持续混乱。例如，部分古镇景区充斥着各种物非所值的消费、不适当的破坏风貌的现代化建设等。而这些可见的问题长期积压，却一直未得以解决。在不同的媒体报道里不难找到民众对旅游体验做出的负面评价——对门票制度、旅游环境等表示不符合预期，但是各景点的旅游人次每年屡攀新高。游客对古镇旅游的态度到底怎样，游客的评判标准和关注热点是什么，哪些因素可能会影响古镇旅游体验，可以运用大数据对游客的态度进行探索，从而为解决这些问题提供参考。

本文结合当前流行的文本挖掘方法，选取了对几个古镇旅游的网络评论进行文本处理。先提取高频词，分析游客的关注热点，对几个旅游景点横向比较的同时，再内部纵向比较游客对哪个旅游属性比较敏感。再利用情感分析SnowNLP组件进行情感分析。运用分析结果探索游客对敏感属性的态度是否影响最后对该景点的总体好感度。最后构建模糊综合评价模型，从而可以得到在大数据的支持下游客对古镇旅游的总体态度。

2 数据描述和方法

2.1 获取数据及数据预处理

使用数据采集器从网站[①]采集所需的评论词条组成数据。本文一共采集了6 647条分别关于甲居藏寨、西江千户苗寨、周庄古镇、黄姚古镇、西塘古镇的评论。对采集的数据用R软件预处理。现在已有很多程序包可以调用，如jieba和Rwordseg等，本文运用的是Rwordseg。Rwordseg是基于Java的程序包，使用它进行分词时要调用Java分词工具Ansj[②]。

首先将采集到的评论进行分词处理，即依据语义将一句话切分成一个个的词，例如，“我也喜欢周庄”分词成：“我”“也”“喜欢”“周庄”。虽然分成了一个个的词，但是句子里有一些没有意义的停词，如例子中的“也”，要将其过滤掉，该过程使用一个含有1 000多个停词的词库，可以实现停词过滤。分词这一步尤为重要，分词效果的好坏直接影响了后面分析的准确性。依据Rwordseg包里自带的词典，文本被分为一个个的词。为了提高准确度，还添加了搜狗词库里针对旅游这一话题的词。为了提取有用的信息，进一步过滤掉一个字的词。最后统计5个景点的评论词频。截取排序前20的词，见**表1**。

依据统计的各景点的词频制作词云图，让统计结果更加直观，便于寻找主要属性特征。鉴于一些高词频结果与研究目的无关（如关于周庄的词频统计里，“周庄”一词出现了1 894次，显然，这是对周庄的评论，自然“周庄”是高频词）。筛选掉无意义的高频词，制作得到词云图，如**图1~图5**所示。

① http://www.mafengwo.cn

② Ansj是基于中国科学院计算技术研究所ICTCALS（Institute of Computing Technology, Chinese Lexical Analysis System）中文分词算法的开源工具

③ https://github.com/isnowfy/snownlp

表1 词频

词序	甲居藏寨		西江千户苗寨		周庄古镇		黄姚古镇		西塘古镇	
	词	频数	词	频数	词	频数	词	频数	词	频数
1	中国	122	苗寨	201	周庄	1 894	古镇	711	古镇	1 662
2	乡村	94	西江	113	古镇	1 064	一个	156	江南	774
3	一个	89	晚上	73	水乡	852	感觉	110	感觉	605
4	门票	88	夜景	66	江南	777	地方	102	水乡	580
5	特色	84	看到	61	感觉	346	门票	94	一个	545
6	地方	81	全景	59	一个	336	不错	83	地方	458
7	藏族	64	千户	51	商业化	329	安静	82	酒吧	446
8	漂亮	60	整个	48	门票	300	喜欢	79	喜欢	432
9	美丽	60	风景	46	小桥流水	280	小镇	73	晚上	432
10	民居	57	地方	45	地方	278	客栈	69	乌镇	412
11	感觉	57	拍照	42	不错	271	商业化	68	门票	351
12	风景	57	不错	31	万三	250	房子	64	不错	318
13	比较	55	台上	30	比较	231	晚上	62	商业化	308
14	寨子	54	景色	28	商业	220	特色	61	比较	307
15	整个	51	感觉	28	晚上	217	宁静	61	小镇	283
16	藏式	48	白天	28	中国	211	比较	54	生活	264
17	梨花	47	漂亮	27	第一	200	值得	53	烟雨	250
18	建筑	47	观光车	26	景点	198	石板	53	特色	237
19	景色	43			气息	192	里面	51	觉得	226
20	山坡	42			喜欢	187	游客	50	景区	225

2.2 总结热点词及情感分析

基于词频统计结果，分析词频表，合并同义词，再组成话题热词。例如，5个景点都有对景区门票、景区内物价的评论。把该类关于消费的词组合并，提取出一个属性，叫做“价格制度”。其中，黄姚古镇的“门票”“学生证”“免费”等词的词频数之和为119，可计算价格制度词频数及该话题所占比重。最终得到的属性见**表2**。

本文还运用了Python对文本进行情感分析，其中运用了SnowNLP[③]组件。SnowNLP是一个利用Python编写的类库，主要针对中文分词等文本处理，优点是自带了一些训练好的字典，用起来比较方便，准确度也高。Python软件运行情感分析的SnowNLP库。其计算情感指数的原

图1 甲居藏寨词云图

图2 西江千户苗寨词云图

表 2 特征属性的词频及比重

属性	甲居藏寨		西江千户苗寨		周庄古镇		黄姚古镇		西塘古镇	
	词频	比重	词频	比重	词频	比重	词频	比重	词频	比重
价格制度	103	0.089 57	19	0.022 25	362	0.078 7	119	0.092 82	448	0.087 53
商业化	37	0.032 17	13	0.015 22	656	0.142 61	161	0.125 59	641	0.125 24
历史人文	376	0.326 96	217	0.254 1	2 320	0.504 35	489	0.381 44	2054	0.401 33
自然风光	224	0.194 78	244	0.285 71	474	0.103 04	104	0.081 12	1016	0.198 52
特色	410	0.356 52	361	0.422 72	788	0.171 3	409	0.319 03	959	0.187 38

图 3 周庄古镇词云图

图 4 黄姚古镇词云图

图 5 西塘古镇词云图

理是：首先，规定情感为两类——积极情感和消极情感。根据贝叶斯定理，文本属于积极类（C）的概率是由类别的概率P(C)乘以每个文本在类别C中的条件概率得来的：

$$P = P(\mathrm{C}) \cdot \Pi_i^n P(d_i \mid \mathrm{C}) = P(\mathrm{C}) \cdot \Pi_i^n \frac{count(d_i, \mathrm{C})}{T_C} \quad (1)$$

其中，$count(d_i,\mathrm{C})$表示词d_i在C中出现的次数，T_C是C类别的词总数，n是进行分类的文本中词组的数量。运用该组件得到的评分结果取值范围是（0,1）。

3 研究结果及数据展示

3.1 热点话题的构成情况

从**图6**可以得到直观的结论。纵向来看，对于每个古镇，其历史文化被提到的次数最多。可以看出，游客选择到古镇旅游的目的是明确的，这一类旅游项目提供的主要是历史人文的氛围。在5个景区的属性占比里，自然风光占总关注度比例最少（除了西江千户苗寨），自然风景也不是该类旅游项目的重点。研究对象中，周庄的历史人文关注度最高。西江千户苗寨的民族特色占比最多，其次是自然风光。而深入了解可知，西江千户苗寨主打的苗族特色鲜明，并且房屋依山而建，加上山顶的观景台，人文与自然风景相结合，因此游客对风景的关注度也比较高。甲居藏寨的关注热

点是独具的藏族民俗文化和建筑风格，即民族特色，包括特有的少数民族的习俗、特有的人文风貌。西塘古镇的话题结构与周庄古镇相似，也有江南水乡的美誉。

横向来看，最具特色的是西江千户苗寨和甲居藏寨，自然风景占比最高的是西江千户苗寨，而最具人文历史的景区是周庄古镇。每个景点都有不同程度的商业化，其中较为严重的是黄姚古镇、周庄古镇、西塘古镇这3个景区。可以从过去景区规划和发展里看出，这些地方已经被开发得比较好了，但是过度商业化造成大多古镇雷同，失去自身特色。西江千户苗寨和甲居藏寨商业化较轻，这与地理位置有关，一般来说偏远村落能更好地保持其原生态。在价格消费方面，西江千户苗寨的关注度是最少的。经查评论，黄姚古镇、周庄古镇和西塘古镇的门票关注度是价格属性里最高的，可见人们对景区门票等收费制度也是比较看重的。

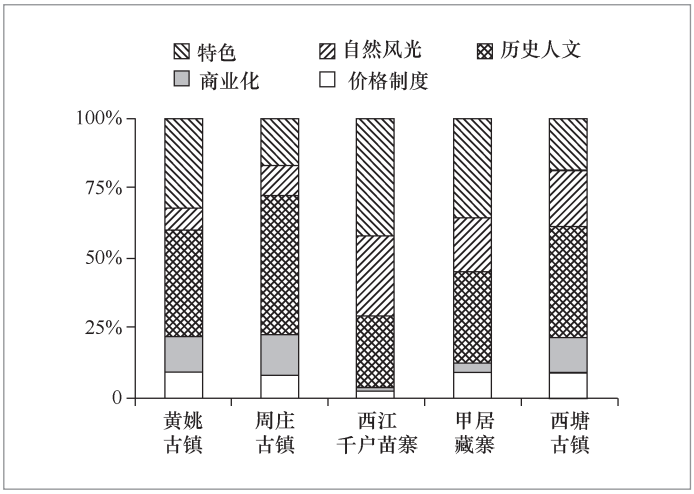

图 6 各属性比重

3.2 收费制度对评分的影响

单因素方差分析用来分析一个因素是否明显造成两组数据的不同。其中，实验结果*P*值是决定是否接受原假设的关键阈值，它体现了实验组别之间差别的显著性。如果*P*值小于0.05，就有统计意义；如果大于0.05，说明所有组别都没有差别，即这个因素不对样本造成影响。在景点属性里，评论展现了人们对景区的价格制度具有较多的负面评价，而门票价格是价格制度里的主要组成。为证实人们确实不满意景区的收费制度，以门票为因素，假设其不对评论结果造成影响，进行单因素方差分析。若实验结果*P*值小于0.05，拒绝原假设，并认为门票制度确实对评价结果造成显著差别。

除了西江千户苗寨的评论对门票关注度比较低，其他几个景点都很高。因此分析各景点门票是否对最终体验好感度造成明显差异是有必要的。基于情感分析输出结果，对各古镇门票进行单因素方差分析。先将包含“门票”一词的评论和没有该词的评论分开，单独构成数据，再进行单因素方差分析，分析结果见**表3**。本节没有包括西江千户苗寨的方差分析结果，原因是只有该古镇的实验结果*P*值大于0.05，“门票”因素才不造成显著影响，无需进一步讨论。观察**表3**的实验结果可知，门票确实对甲居藏寨、周庄古镇、西塘古镇、黄姚古镇的游客的评分造成显著影响。并且包含门票的评分均值都低于不包含门票的评分

表 3 单因素方差分析结果

项目	甲居藏寨	周庄古镇	黄姚古镇	西塘古镇
包含“门票”评分均值	0.694 822 95	0.525 235 18	0.473 015 84	0.390 611 75
不包含“门票”评分均值	0.857 206 41	0.693 428 18	0.699 982 77	0.738 293 33
显著性（*P*值）	0	0	0.000 1	0

均值。查看包括“门票”的原评论，发现不少人表示门票涨价，但景区内容质量却没变。甚至有商贩在不该收取门票的地方收取费用，还出现一些欺骗案例。可见古镇的旅游市场秩序欠缺严格的管理，导致游客对这方面有强烈的负面情绪。

4 古镇模糊综合评价模型

本节将对整个古典旅游进行最终评分。每个古镇由于地理位置、开发程度和所在地的经济发展等因素不同，游客的好感度也会受影响。有的省份注重旅游业的发展，旅游市场规划和管理得较好，且服务业的发达程度也会影响人们的满意度。为了对古镇的旅游作出最终评分，构建了一个模糊综合评分模型。由于旅游业属于第三产业，为了方便获取数据，笔者取各景区所在地的第三产业占比作为因子分析的数据。目的是依据各地的第三产业的重要性决定各景区的权重。把5个景点当作5个不同的属性，旅游业发展规划良好的地区应赋予较大权重。

4.1 第三产业占比的因子分析

因子分析的基本目的就是用少数几个因子描述许多指标或因素之间的联系，即将比较密切的几个变量归在同一类中，每一类变量就成为一个因子，以较少的几个因子反映原资料的大部分信息。每个因子中，因子载荷越大，说明该因子对结果的贡献越大。运用因子分析可以得到景区所属地区的第三产业的贡献率，即可得到各地区的旅游业权重。所谓权重，是指某指标在整体评价中的相对重要程度。权重越大，则该指标的重要性越高，对整体的影响就越高。本文选取每个景区所属省份从1990年到2015年第三产业的贡献率进行因子分析。首先进行KMO（Kaiser-Meyer-Olkin）检验，检验结果见表4。KMO是统计产品与服务解决方案（statistical product and service solutions，SPSS）提供的用于判断原始变量是否适合进行因子分析的统计检验方法之一。*KMO*值的大小可以决定是否有进一步分析的必要。若*KMO*值大于0.5，则因子分析可以进行。由**表4**可知，*KMO*检验结果为0.735 9，大于0.5，数据可以做因子分析，v_1~v_5表示5个景区所属省份的第三产业占比。

选取特征值大于1且贡献率大于80%的因子。只有一个特征值大于1，为3.320 32，见**表5**。后面4个因子可以省去，选取特征值最大的公共因子。由因子载荷得到归一化处理的权重*w*，见**表6**。

$$w=(0.239\ 7,0.244\ 2,0.096\ 2,0.183\ 9,0.236\ 0) \quad (2)$$

4.2 综合评价模型构建

为了得到总体游客的态度，不能单纯地计算评分结果的均值。原因是游客的评价通常不是单一的积极或消极这么简单。机器计算结果是基于字面表达上的积极词和消极词得来的。但是具体评判分值只依据一段文字表达不够准确。建立模糊评价模型可以包容情感分析带来的文字到数值转化的误差。而且由于模糊的方法更接近东方人的思维习惯，因此更适合对社会经济系统问题进行评价。

对5个景点的情感评分分别以条件$0\leqslant q\leqslant 0.3$,$0.3\leqslant q\leqslant 0.6$, $0.6\leqslant q\leqslant 1$分为3组，表示态度消极、中肯、积极，并计算各组占比，计算结果见**表7**。把原来0~1的评分结果转化成1~5分的评分标准。转化为5分制后，个人的情感倾向更加分明。

得到比重矩阵R：

$$R=\begin{pmatrix}0.215\,5 & 0.117\,4 & 0.667\,1\\ 0.188\,0 & 0.093\,5 & 0.718\,5\\ 0.244\,9 & 0.100\,2 & 0.654\,9\\ 0.109\,7 & 0.055\,8 & 0.834\,6\\ 0.085\,2 & 0.068\,9 & 0.845\,9\end{pmatrix} \quad (3)$$

将消极、中肯、积极分别赋予分值1、3、5。计算最终评分：

$$\begin{aligned}Q&=w\cdot R\cdot\begin{pmatrix}1\\3\\5\end{pmatrix}\\&=(0.239\,7,0.244\,2,0.096\,2,0.183\,9,0.236\,0)\\&\begin{pmatrix}0.215\,5 & 0.117\,4 & 0.667\,1\\ 0.188\,0 & 0.093\,5 & 0.718\,5\\ 0.244\,9 & 0.100\,2 & 0.654\,9\\ 0.109\,7 & 0.055\,8 & 0.834\,6\\ 0.085\,2 & 0.068\,9 & 0.845\,9\end{pmatrix}\begin{pmatrix}1\\3\\5\end{pmatrix}\\&=(0.161\,4,0.087\,1,0.751\,5)\begin{pmatrix}1\\3\\5\end{pmatrix}=4.180\,2\end{aligned} \quad (4)$$

最终得到的评分Q为4.180 2，接近5分满分。可以看出，虽然游客对一些商业化和不合理收费等有明显负面情绪，但总的来说对古镇旅游体验是比较满意的。该模型可以运用到国内更多数量的古镇旅游的评分中。

5 结束语

通过对网络文本数据进行挖掘和统计分析以及进一步构建模糊综合评价模型，得出游客对古镇旅游态度及相关因素的结论如下。

- 游客选择古镇旅游，关注度最高的是其文化价值，其次是各个古镇的特色及环境。随着景区的开发，商业气息趋严重。而伴随着商业化的同时，每个古镇的特色会削弱，对游客的体验造成负面影响。游客对景区里的消费也十分敏感，在黄姚古

表4 KMO检验

变量	*KMO*
v_1	0.515 7
v_2	0.814 8
v_3	0.692 9
v_4	0.737 3
v_5	0.803 8
总体	0.735 9

表5 提取因子

因子	特征值	差异	贡献率	累计贡献率
因子1	3.320 32	2.474 74	0.832 4	0.832 4
因子2	0.845 58	0.839 00	0.212 0	1.044 3
因子3	0.006 58	0.037 94	0.001 6	1.046 0
因子4	−0.031 36	0.120 75	−0.007 9	1.038 1
因子5	−0.152 11	/	−0.038 1	1.000 0

表6 因子载荷

变量	因子1	归一处理
周庄古镇	0.940 1	0.239 705 244 906 805
西塘古镇	0.957 9	0.244 243 861 393 712
黄姚古镇	0.377 1	0.096 152 375 124 302
甲居藏寨	0.721 2	0.183 890 461 256 024
西江千户苗寨	0.925 6	0.236 008 057 319 157
总数	3.921 9	1

表7 情感分组

景点	消极		中肯		积极	
	评论数/条	比重	评论数/条	比重	评论数/条	比重
周庄古镇	494	0.215 5	269	0.117 4	1 529	0.667 1
西塘古镇	559	0.188 0	278	0.093 5	2 136	0.718 5
黄姚古镇	132	0.244 9	54	0.100 2	353	0.654 9
甲居藏寨	59	0.109 7	30	0.055 8	449	0.834 6
西江千户苗寨	26	0.085 2	21	0.068 9	258	0.845 9

镇、西塘古镇、周庄古镇，游客对商业化感受比重占到了12%以上，应该引起有关部门的重视，予以管束，在开发特色旅游的同时，保护景区的环境，营造良好的历史文化氛围。

● 除了不适当的商业化影响游客对古镇旅游的态度，消费制度也是关注热点。本文研究了游客对价格制度的态度及其影响。结论基于对门票价格的分析，游客确实对收费制度敏感，且对收费敏感的游客均持负面评价。在选取的5个研究对象里，只有一个古镇的游客表示门票等收费合理。大部分游客认为景区内物非所值，甚至抱怨有不透明收费的现象。这种充斥着消费气息的古镇旅游项目不利于长久发展。景区应结合自身的服务设施等调节景区内消费价格，整改冗余的收费项目。有关部门应予以管制，规范旅游市场。

● 结合地区第三产业的比重，得到游客对古镇旅游的模糊综合评价。从最终评分结果来看，游客的总体态度是积极的。这也解释了为什么游客在对消费制度和商业化如此敏感的情况下，古镇旅游项目每年接待人次屡创新高。游客对负面影响因素有明显的感知，但是这并没有打击游客的积极性。

参考文献:

[1] 李爽，周璇玲，丁瑜，等. 大陆居民赴台湾旅游体验感知研究[J]. 旅游论坛，2015(8)：7-20.
LI S, ZHOU X L, DING Y, et al. Research on experience perception of mainland residents to Taiwan travel[J]. Tourism Forum, 2015(8): 7-20.

[2] 王耀斌，杨玲，孙传玲，等. 基于文本挖掘的典型旅游网站旅游分享比较研究[J]. 资源开发与市场，2017(1)：100-104.
WANG Y B, YANG L, SUN C L, et al. Comparative research on travel sharing of typical travel website based on text mining—taking Gansu province as an example[J]. Resource Development & Market, 2017(1): 100-104.

[3] 刘萌玥，陈效萱，吴建伟，等. 旅游景区网络舆情指标体系构建[J]. 资源开发与市场，2017(1)：80-84.
LIU M Y, CHEN X X, WU J W, et al. Research on construction of index system of tourist attraction network public opinion[J]. Resource Development & Market, 2017(1): 80-84.

[4] 魏慧玲. 文本情感分析在产品评论中的应用研究[D]. 北京：北京交通大学，2014.
WEI H L. Application research of text sentiment analysis in product reviews[D]. Beijing: Beijing Jiaotong University, 2014.

[5] 林园园，战洪飞，余军合，等. 基于产品评论的消费者情感波动分析模型构建及实证研究[J]. 现代图书情报技术，2016(11)：44-53.
LIN Y Y, ZHAN H F, YU J H, et al. Using product reviews to analyze sentiment fluctuation of consumer[J]. New Technology of Library and Information Service, 2016(11): 44-53.

[6] 陈苗，刘超，庄俊玲，等. 基于文本挖掘的临床带教评价分析[J]. 中国卫生统计，2017(34)：59-64.
CHEN M, LIU C, ZHUANG J L, et al. Analysis of commands for clinical teaching based on text mining[J]. Chinese Journal of Health Statistics, 2017(34): 59-64.

[7] 陈顾远，王超，刘富强，等. 一种基于微博数据的公众环境污染情感指数估算法[C]//2016全国环境信息技术与应用交流大会暨中国环境科学学会环境信息化分会年会，2016年11月24日，北京，中国. 北京：环境保护部环境发展中心，2016：474-481.
CHEN G Y, WANG C, LIU F Q, et al. Estimate of public environment-emotional index based on micro-blog data[C]//2016 National Conference on Environmental Information Technology and Application and Annual Meeting of Environmental Information of China Environmental

Science Society, November 24, 2016, Beijing, China. Beijing: Environment Development Centre of Ministry of Environmental Protection, 2016: 474-481.

[8] 余维军. 基于语义情感信息的评论有用性研究[D]. 重庆: 重庆大学, 2016.
YU W J. Research on review helpfulness based on semantic and sentiment information[D]. Chongqing: Chongqing University, 2016.

[9] 耿铭泽. 基于旅游博客的旅游目的地游客满意度评价研究——以海南省为例[D]. 海口: 海南大学, 2010.
GENG M Z. A study on the tourist satisfaction degree of tourism destination by travel blogs-a case study of Hainan province[D]. Haikou: Hainan University, 2010.

作者简介

范珈瑜(1995-), 女, 山东财经大学统计学专业本科生, 主要研究方向为应用统计学、大数据应用。

收稿日期: 2017-05-22